Educational Administration: Theory and Practice
2024, 30(5), 2174-2181
ISSN:2148-2403
https://kuey.net/

Research Article

Performance Analysis For Detection And Classification Of Lung Cancer Using Machine Learning Approaches

K. Ramkumar[1*], M. Natarajan[2]

[1*]Research Scholar, Department of Computer and Information Science, Faculty of Science, Annamalai University, Annamalainagar, Tamil Nadu, India. Email: ramau13@gmail.com
[2]Department of Computer and Information Science, Faculty of Science, Annamalai University, Annamalainagar, Tamil Nadu, India
Email: mind.2004@gmail.com

Citation: K. Ramkumar, et al (2024), Performance Analysis For Detection And Classification Of Lung Cancer Using Machine Learning Approaches, *Educational Administration: Theory and Practice*, *30*(5), 2174-2181
Doi: 10.53555/kuey.v30i5.3255

ARTICLE INFO

ABSTRACT

Chest X-ray is often the initial imaging test used to detect abnormalities in the lungs, it's not usually sufficient for confirming a diagnosis of lung cancer. Instead, additional imaging tests like CT scans, MRI scans, or PET scans may be needed for a more detailed evaluation. These tests can provide more information about the size, location, and spread of any suspicious areas in the lungs, helping doctors to make a more accurate diagnosis. Automated lung cancer classification is one of the important tasks due to the different mechanisms used for imaging the lungs of patients. Using training and testing methodology, machine learning approaches to identify lung cancer have shown excellent detection and classification potential. In this paper, we have demonstrated an effective approach for detecting and classifying lung cancer-related CT scan images using image processing techniques, and then further supervised machine learning algorithms are used for their classification of lung cancer. We have extracted texture features along with statistical features and supplied various extracted features to classifiers using three different classifiers known as the k-nearest neighbors' classifier, support vector machine classifier, and random forest classifier. Based on the numerical illustrations, the best results are mentioned with high accuracy.

Keywords Lung cancer, Image processing, Image classification, Machine learning and Test Statistics

1 Introduction

According to study, nineteen different types of cancer can affect a healthy person. Among all these cancers, lung cancer has the highest mortality rate. It is estimated that about 1.7 million people die annually due to this disease. Principle cause of lung cancer has been attributed to smoking which accounts around 80% of total lung cancer cases worldwide. In its initial stage, it is difficult to detect the lung cancer. As per findings, about 25% of people who were diagnosed with lung cancer in its initial stage experienced no symptoms at all. Unlike other cancers, lung cancer cannot be seen with naked eye and its symptoms are often masked with other disease symptoms such as bronchitis, asthma, and coughing.

Today, cancer has become one of the most common causes of death among youths worldwide. Lung cancer, breast cancer, stomach cancer, and prostate cancer are one of the most frequently diagnosed cancers among men and women that lead severe complexities or in many cases to death, if not detected at early stage. Cancer in human body represents the abnormal growth of cell. Lung cancer is usually identified when an X-ray or CT scan of patient's chest is performed for any another good reason [2]. The rest of 75% people are diagnosed when they experience or develop some sort of symptoms. These symptoms may arise due to direct effects of the primary tumor or due to effects of cancer that has spread to other parts of the body (metastases) or due to disturbances of hormones, blood, or other systems. Lung cancer usually spreads toward the center of the chest cavity; this is because the natural flow of lymph which is outward from the lungs and inward toward the center of the chest [3]. When cancer sets in, a single cell or a bunch of cells abruptly start multiplying in an uncontrolled and disorganized way, which if not stopped can lead to formation of lumps or tumors. Any tumor can be broadly classified into two categories benign and malignant.

In contrast to benign tumor, malignant tumor penetrates nearby body cells and can spread to other body parts with its growth. Usually, benign tumors are not considered very dangerous but can become dangerous if they develop vital structures required for their growth such as blood vessels or nerves. As the discussed pattern is followed, the tumor is said to be invasive and radiologists and other healthcare professionals use the various imaging techniques for its detection.
Computed tomography (CT) scans of lungs. Figure 1 depicts a general methodology of lung cancer detection system that consists of five basic stages. The first stage is of image acquisition that represents the collection set of images related to body part. For this work, we have obtained DICOM CT scan images of lungs from an online database [4]. This database was part of the 2015 SPIE Medical Imaging Conference, SPIE with the support of American Association of Physicists in Medicine (AAPM), and the National Cancer Institute (NCI) conducted a Grand Challenge on quantitative image analysis methods for the diagnostic classification of malignant and benign lung nodules.

2 Related work

In recent years many machine learning approaches, especially neural networks, have been widely used for detection of lung cancer using medical images. Many of the proposed
approaches have achieved high accuracy rate [5].

2.1 Image processing techniques

Dwivedi et al. [6] have proposed a image pre-processing method known as contrast limited adaptive histogram equalization (CLAHE). They have used gray-level co- occurrence matrix (GLCM) for extracting the image features, which also gives the information regarding the position of those pixels which have similar values of gray level.
A GLCM can contain a variety of statistical features that can be extracted from the matrix for analysis purpose. Authors have used automatic feature selection algorithms for determining the best features. Qian et al. [7] have conducted a study to predict the near-term risk of developing breast cancer by using the image dataset of mammograms. Proposed scheme is designed around the four image processing modules like image pre-processing, segmentation, feature extraction, and classification to compute image feature asymmetry. Chaudhary and Singh [8] have processed the image by applying (i) image pre-processing and (ii) feature extraction methods. Image pre-processing step consists of two major segments: image enhancement and image segmentation.
Pratap and Chauhan [9] have applied watershed techniques to perform image segmentation. Bhusri et al. [10] have used law's of feature extraction in order to extract features from the region of interest. Many other image processing techniques for image enhancement, pre-processing, segmentation, and feature extraction have been suggested by [5]. Kuruvilla and Gunavathi [11] have used simple Otsu's method for image segmentation, along with morphological opening method, with periodic line as the structuring element of fixed in size. In order to overcome this issue, they have suggested that morphological operations must be performed to fill in the gaps left after applying thresholding method on the grayscale CT scan images.

Fig. 1 Steps for detection of lung cancer

2.2 Classification using machine learning approaches

Neural networks form the basis for many machine learning- based approaches. It is the task for classifying an image into various classes based on the labels of the input training dataset. There are various machine learning algorithms which can be used for the task of image classification and can be classified into two categories known as supervised and unsupervised learning approaches. Mitra and Pal [12] and Amato et al. [13] have obtained the similar results. Karabatak and Ince [14] have used a combination of association rules and neural network to provide efficient computer-aided diagnostic system. The use of association rules reduces the dimensions of feature vector, without greatly compromising the accuracy of the overall system. Dwivedi et al. [6] have used multinomial multivariate Bayesian references for classification of image obtained after image processing stage, into normal (non-cancerous) and abnormal (cancerous) images. Adi et al. [15] have developed a system based on digital image processing techniques for identification of cancer cells through the stages of feature extraction using GLCM and classification using a naive Bayes algorithm.
In their results they have achieved the accuracy of 88.57% in detecting the lung cancer. Joachims [16] has presented an improved algorithm for training support vector machine (SVM) on largescale datasets and problems and described its effective way of implementation in SVM. Joachims has also introduced the technique for shrinking the size of problem during its optimization. Tidke and Chakkarwar [17] have presented a CAD system for early detection of lung cancer from CT images and to classify whether tumor is benign or malignant. They have presented the five-stage model and used GLCM for textural feature extraction and SVM

classifier for image classification. In their experiments they have used small size of dataset which consists of only 25 JPEG images and obtained results show 96% accuracy using SVM classifier. Touw et al. [18] have listed the key features of random forest and mentioned that this algorithm is the most widely used in life sciences for both regression and classification of tasks, for example the prediction of disease state of patients. It also allows the extraction of additional relevant knowledge from omics data. Shi et al. [19] have presented a random forest clustering strategy for tumor profiling based on tissue microarray data. They have used this method to detect and analyze the renal cell carcinoma a type of kidney cancer in adults. Ramos-Gonzalez et al. [20] have proposed a novel case-based reasoning framework for diagnosis of lung cancer subtypes. They have used gradient boosted regression trees (GBRT) feature selection method to achieve high predictive accuracy. In their experiments they have used knearest neighbors (kNN), naive Bayes classifier (NB) and support vector machine (SVM) methods.

3 Methodology

For obtaining better results we have broadly divided the various stages of Fig. 1 into two different categories known as image analysis and image classification.

3.1 Image Analysis

Image analysis is the process of working on images in order to improve the image quality for human readability and for removing any noise present in images for its efficient classification. Image pre-processing and segmentation, the first step of image analysis is image pre-processing. The acquired image is converted into a grayscale image as shown in Fig. 2; afterward, we apply image denoising methods to remove the noise in the image. Here, we have considered three image denoising methods known as median blur, Gaussian blur, and bilateral blur, as shown in Fig. 3a–c, respectively. After analyzing the results it is found that Gaussian blur outperforms other methods.

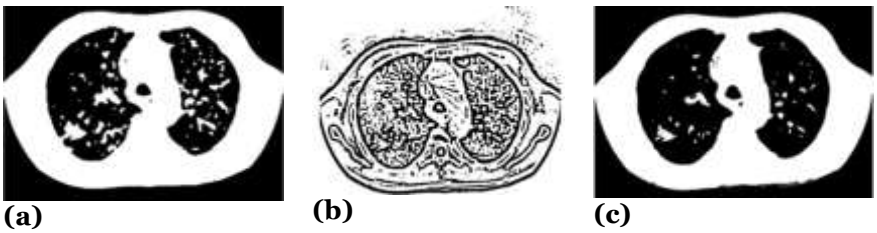
Once image denoising is done, we apply thresholding methods for converting the grayscale image into a binary image. A binary image consists of only two pixel values either "0" or "1," whereas in a grayscale image each pixel can acquire any value between 0 and 255. In this paper, we have compared three thresholding methods known as global thresholding, Otsu's method with adaptive mean thresholding, and Otsu's method with adaptive Gaussian thresholding. While comparing these three thresholding methods it is found out that Otsu's method with adaptive Gaussian thresholding provides the suitable results for further analysis.

Here, Fig. 4a–c represents the various results of these thresholding methods, respectively, on the image obtained after applying Gaussian blur operation. After thresholding, morphological opening operation is applied on the image to fill in the gaps left after thresholding. Morphological operations are some of the simplest operations that can be performed on an image based on its shape.

Fig. 2 Input image

(a) **(b)** **(c)**

Fig. 3 Applying image denoising methods on input image a) median blur, b) Gaussian blur and c) bilateral blur



(a) **(b)** **(c)**

Fig. 4 Image processing after Gaussian blur operation a) global thresholding, b) Otsu's method with adaptive mean thresholding, and c) Otsu's method with adaptive Gaussian thresholding.

GLCM features GLCM has a total of fourteen different features, but among them the most useful features are: contrast, dissimilarity, homogeneity, correlation, angular second moment (ASM), and energy are considered in this paper.
Statistical features From the region of interest we have extracted six statistical parameters, namely standard deviation, skewness, kurtosis, fifth and sixth central moments, root-mean-square, and mean.
Mean It is the average of all pixel intensity values and can be expressed as [11]:

$$\mu = \frac{1}{MN}\sum_{i=0}^{M}\sum_{j=0}^{N} p(i,j) \qquad \ldots (1)$$

where p(i, j) is the value of pixel intensity at the point (i, j), and M × N is the size of the image.
Standard Deviation: is the estimation of mean square deviation of the gray pixel value p(i, j) from its mean value μ [11].

$$\mu = \sqrt{\frac{1}{MN}\sum_{i=0}^{M}\sum_{j=0}^{N} p(i,j) - \mu)^2} \qquad \ldots (2)$$

Skewness characterizes the degree of asymmetry of a pixel distribution in the specified window around its mean. It is a pure number, and it characterizes only the shape of the distribution [11].

$$S = \frac{1}{MN}\sum_{i=0}^{M}\sum_{j=0}^{N}\left[\frac{p(i,j)-\mu}{\sigma}\right]^3 \qquad \ldots (3)$$

Root-mean-square error It measures the error between the predicted value and the known value. Root-mean-square error can be calculated as:

$$rmse = \sqrt{\frac{p(i,j)^2}{M \times N}} \qquad \ldots (4)$$

Dissimilarity is defined as the sum of pixel values where p (i, j) is the absolute difference between I and j and can be represented as:

$$Dissimilarity = \sum_{i=0}^{M}\sum_{j=0}^{N} p(i,j)\,|i-j| \qquad \ldots (5)$$

Fig. 5 Image obtained after morphological opening operation

This results in high value of IDM for homogeneous images and relatively low IDM value for inhomogeneous images. Correlation is a measure of gray-level linear dependence between the pixels at the specified positions relative to each other and can be expressed as:

$$Correlation = \sum_{i=0}^{M-1}\sum_{j=0}^{N-1}\frac{\{i \times j\} \times p(i,j) - \{\mu_x \times \mu_y\}}{\sigma_x \times \sigma_y} \qquad \ldots (6)$$

Image classification It is the task of classifying the input image into two classes where "0" represents benign tumor, and '1' represents malignant tumor. If the whole available dataset is used as the training set, the classifier will simply memorize the dataset and provides 100% accuracy on the dataset, but as soon as new images are included in the dataset the classifier performs poorly.

The training set consists of 70% of the total images in the dataset, whereas both the test set and the validation set each consist of 15% of the total images in the dataset. Here, we have used four supervised machine learning algorithms to perform classification tasks, known as K-nearest neighbors (KNN) classifier, support vector machine (SVM) classifier, and random forest classifier.

4 Proposed Algorithm

A new image processing technique has been proposed for extracting various features from the images of lung cancer dataset. These features were then used by various supervised machine learning algorithms to detect and classify cancerous mass present. The proposed image pre-processing pipeline helps the machine learning algorithms to better predict the presence of cancerous mass. The results presented in Fig. 6 show the classification of accuracy of each of the seven machine learning algorithms along with dependency of each of the classifiers on the hand-crafted features. As presented in Algorithm 1, input to the proposed algorithm consists of three sets, namely cancerous image set (I), label set (L), and position set P_X, P_Y. The cancerous image set consists of CT scan images of patient's lungs, the label set contains labels for each image in the set "I" marked as either benign (0) or malignant (1), and the position set contains the (x, y) coordinates of cancerous mass present in each image of the set "I."

5. Performance evaluation

The dataset used included 512 × 512 pixels images, categorized over 2 classes—benign and malignant. The total size of the dataset is ≈ 11:2 GB. For each image containing the cancerous portion region of interest was extracted which was a generalized area of 130×130 pixels. The center for this area was obtained from the CSV file provided with the dataset. This complete setup was run using OpenCV for image processing tasks, along with scikitlearn used as a machine learning library written in python. Here, Table 1 represents the details of the experimental setup. Once ROI is extracted from each image, all the above-mentioned seven machine learning algorithms are evaluated at the configured experimental setup on the basis of the following four parameters which in turn is calculated using the confusion matrix.

Accuracy is a statistical measure of how well a classifier correctly identifies or excludes a condition. The accuracy is the percentage of true results (both true positive and true negative) in the given dataset [11]. The author [21-22] proposed a technical report, evaluating results of Machine Learning experiments using Recall, Precision, and F-Measure. In the Medical Sciences, Receiver Operating Characteristics (ROC) analysis has been borrowed from signal processing to become a standard for evaluation and standard setting, comparing True Positive Rate and False Positive Rate.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad \ldots (7)$$

F1 score is a measure of classifiers accuracy. It uses both precision and recall to calculate the score and can be represented as:

$$F_1 Score = \frac{Precision.Recall}{Precision + Recall} \quad \ldots (8)$$

Precision measures the total number of positive cases which the algorithm identifies as per the following [11]:

$$Precision = \frac{TP}{TP + FP} \quad \ldots (9)$$

Recall also known as sensitivity and measures the percentage of actual positive cases which the algorithm correctly identifies. That is the percentage of the images containing a benign or malignant nodule correctly classified by the algorithm as benign or malignant.

$$Recall = \frac{TP}{TP + FN} \quad \ldots (10)$$

True positive (TP) Images containing benign/malignant, classified as benign/malignant. False negative (FN) Images containing benign/malignant, classified as malignant/benign. True negative (TN) Images not containing benign/malignant classified as not containing benign/malignant. False positive (FP) Images not containing benign/malignant classified as containing benign/malignant.

As obtained after simulations, figure 6, figure 7 and figure 8 represent accuracy percentage, F1 score, precision, and recall of each of various mentioned machine learning algorithms for the given dataset. All the classifiers

were trained by using the previously discussed GLCM feature set containing all features. Further, performance evaluation using the abovementioned four parameters for measuring the performance of a classifier was then calculated on subset of the total dataset. This was done so that a uniform result can be obtained, without the possibility of overfitting the dataset. From figure 6, figure 7 and figure 8, we can find that the highest classification accuracy percentage of 92.52% was obtained by using KNN classifier.

Here, tables 2, 3, and 4 represent the effect of removing features from the feature set given to each classifier for classification. Also, tables 2, 3, and 4 represents that accuracy is calculated in percentage, whereas F1 score, precision, and recall values are calculated on the scale of 0–1 converted in the form of percentage.

Table 2. Feature removing using KNN classifier

Features	Accuracy	F1-Score	Precision	Recall
Mean	92.52	75.12	76.54	75.86
SD	87.24	52.54	52.12	52.56
Skewness	91.06	51.51	51.79	52.64
RMSE	90.55	50.12	51.85	50.20
Dissimilarity	89.48	50.85	52.66	51.96
Correlation	88.57	53.32	58.12	52.34

Table 3. Feature removing using SVM classifier

Features	Accuracy	F1-Score	Precision	Recall
Mean	58.1245	40.25	39.87	40.24
SD	59.4712	52.57	52.36	51.87
Skewness	59.7845	57.26	57.12	54.12
RMSE	57.5412	59.84	60.21	59.84
Dissimilarity	56.2112	40.12	42.11	40.17
Correlation	56.1244	54.21	55.62	54.21

Table 4. Feature removing using Random Forest classifier

Features	Accuracy	F1-Score	Precision	Recall
Mean	86.2151	85.21	88.24	84.12
SD	85.1123	85.14	90.12	86.12
Skewness	85.1212	93.24	93.11	94.11
RMSE	81.2124	84.19	84.11	86.88
Dissimilarity	82.1414	77.85	78.54	77.87
Correlation	81.2122	72.54	73.88	74.77

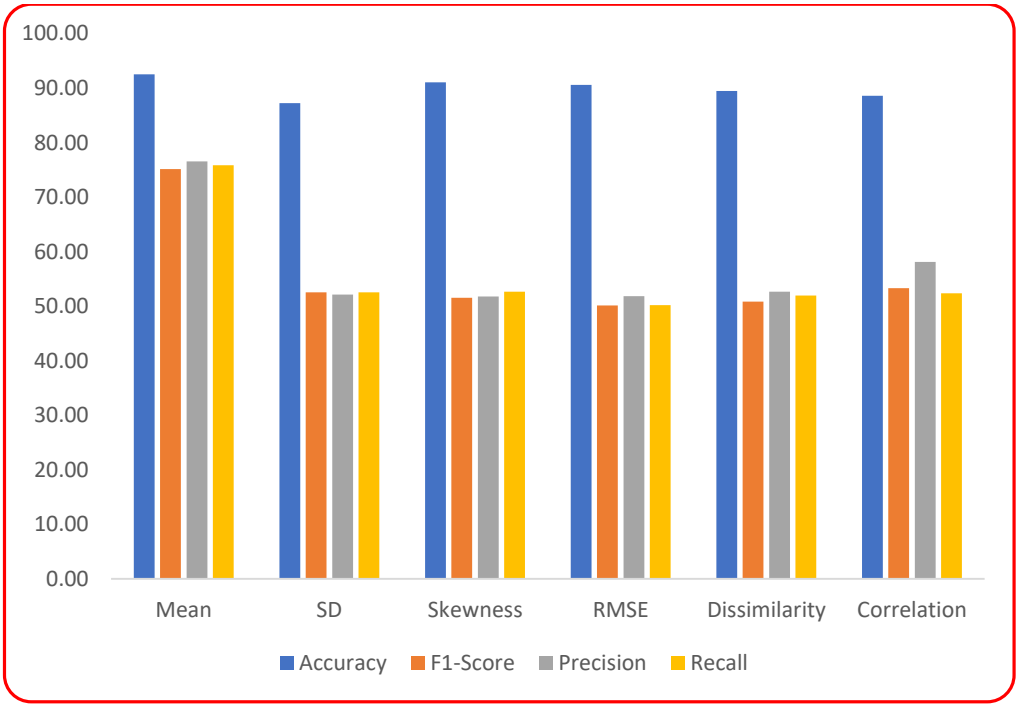


Fig. 6. Performance evaluation using KNN classifier

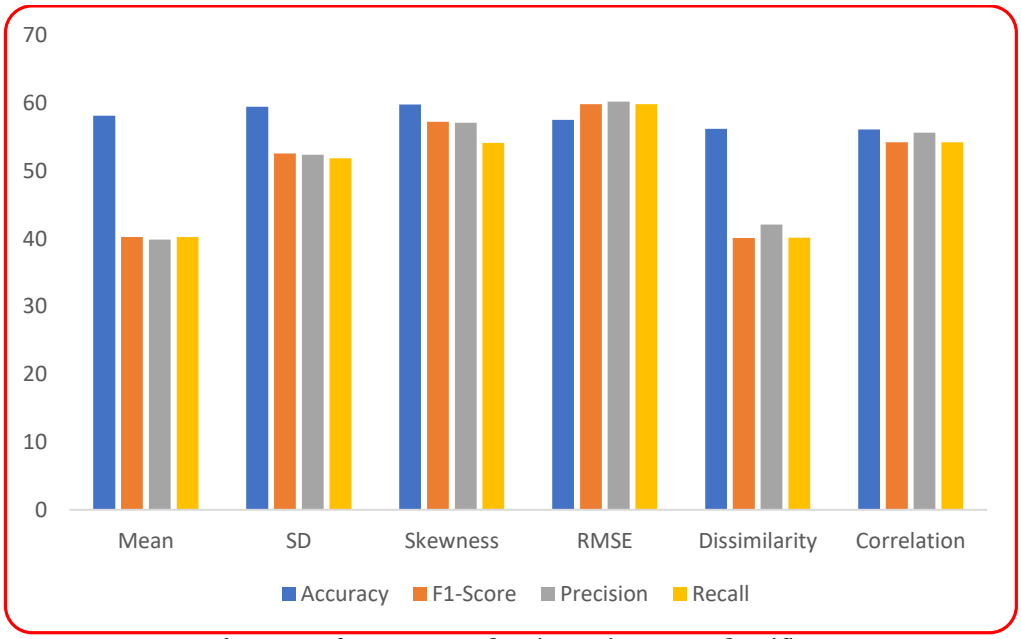


Fig. 7. Performance evaluation using SVM classifier

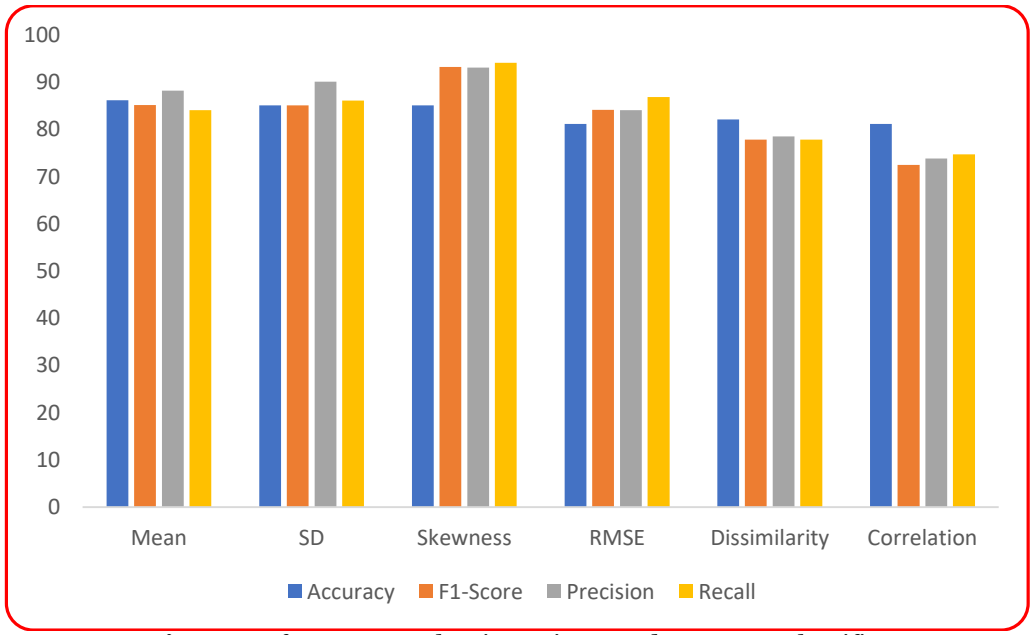


Fig. 8. Performance evaluation using Random Forest classifier

6 Conclusion

In this paper, we have applied image processing and machine learning approaches for detection and classification of lung cancer. The techniques have been categorized and implemented in five different stages known as image acquisition, image pre-processing and segmentation, feature extraction, image classification, and performance evaluation. Simulation results were obtained using four different parameters known as accuracy, F1 score, precision, and recall. Here, obtained results represent that KNN can be applied for detection and classification of lung cancer CT scan images that claims high accuracy.

References

1. Cancer Research UK (2017) Cancer mortality for common cancers. http:// www.cancerresearchuk.org/health-professional/cancer-statistics/mortality/common-ancers-compared. Accessed May 2017
2. Dimililer K, Ugur B, Ever YK (2017) Tumor detection on CT lung images using image enhancement. Online J Sci Technol 7(1):133–138.
3. Al-tarawneh MS (2012) Lung cancer detection using image processing techniques. Leonardo Electron J Pract Technol 20:147–58.

4. Armato III SG, Hadjiiski L, Tourassi GD, Drukker K, Giger ML, Li F, Redmond G, Farahani K, Kirby JS, Clarke LP (2015) SPIEAAPM- NCI Lung nodule classification challenge dataset. The Cancer Imaging Arch. https://doi.org/10.7937/K9/TCIA.2015. UZLSU3FL
5. Gonzalez RC, Woods RE (2002) Digital image processing. Prentice Hall, Upper Saddle River, NJ, pp 797–800.
6. Dwivedi MS, Borse MR, Yametkar MA (2014) Lung cancer detection and classification by using machine learning and multinomial Bayesian. IOSR J Electron Commun Eng (IOSRJECE) 9(1):69–75.
7. Sun W, Zheng B, Lure F, Wu T, Zhang J, Wang BY, Saltzstein EC, Qian W (2014) Prediction of near-term risk of developing breast cancer using computerized features from bilateral mammograms. Comput Med Imaging Graph 38(5):348–357.
8. Chaudhary A, Singh SS (2012) Lung cancer detection on CT images by using image processing. In: Proceedings of 2012 IEEE international conference on computing sciences (ICCS). pp 142–146.
9. Pratap GP, Chauhan RP (2016) Detection of Lung cancer cells using image processing techniques. In: Proceedings of IEEE international conference on power electronics, intelligent control and energy systems (ICPEICES). pp. 1–6.
10. Bhusri S, Jain S, Virmani J (2016) Classification of breast lesions based on laws' feature extraction techniques. In: Proceedings of 3rd international conference on computing for sustainable global development (INDIACom). pp. 1700–1704.
11. Kuruvilla J, Gunavathi K (2014) Lung cancer classification using neural networks for CT images. Comput Methods Programs Biomed 113(1):202–209.
12. Mitra S, Pal SK (1995) Fuzzy multi-layer perceptron, inferencing and rule generation. IEEE Trans Neural Netw 6(1):51–63.
13. Amato F, Lpez A, Pea-Mndez EM, Vahara P, Hampl A, Havel J (2013) Artificial neural networks in medical diagnosis. J Appl Biomed 11:47–58.
14. Karabatak M, Ince MC (2009) An expert system for detection of breast cancer based on association rules and neural network. Expert Syst Appl 36(2):3465–3469.
15. Adi K, Widodo CE, Widodo AP, Gernowo R, Pamungkas A, Syifa RA (2017) Nave Bayes algorithm for lung cancer diagnosis using image processing techniques. Adv Sci Lett 23(3):2296–2298.
16. Joachims T (1998) Making large-scale SVM learning practical (No. 1998, 28). In: Technical Report, SFB 475: Komplexittsreduktion in Multivariaten Datenstrukturen, Universitt Dortmund, pp 1–18.
17. Tidke SP, Chakkarwar VA (2012) Classification of lung tumor using sVM. Int J Comput Eng Res 2(5):1254–1257.
18. Touw WG, Bayjanov JR, Overmars L, Backus L, Boekhorst J, Wels M, van Hijum SA (2012) Data mining in the life sciences with random forest: A walk in the park or lost in the jungle? Brief. Bioinform. 14(3):315–326.
19. Shi T, Seligson D, Belldegrun AS, Palotie A, Horvath S (2005) Tumor classification by tissue microarray profiling: random forest clustering applied to renal cell carcinoma. Mod Pathol 18(4):547–557.
20. Ramos-Gonzlez J, Lpez-Snchez D, Castellanos-Garzn JA, de Paz JF, Corchado JM (2017) A CBR framework with gradient boosting based feature selection for lung cancer subtype classification. Comput Biol Med 86:98–106.
21. Rajesh, P. and Karthikeyan, M., 2017. A comparative study of data mining algorithms for decision tree approaches using weka tool. Advances in Natural and Applied Sciences, 11(9), pp.230-243.
22. Rajesh, P., Karthikeyan, M., Santhosh Kumar, B. and Mohamed Parvees, M.Y., 2019. Comparative study of decision tree approaches in data mining using chronic disease indicators (CDI) data. Journal of Computational and Theoretical Nanoscience, 16(4), pp.1472-1477.